TRANSCRIPTION OF BROADCAST NEWS SHOWS WITH THE IBM LARGE VOCABULARY SPEECH RECOGNITION SYSTEM

Raimo Bakis, Scott Chen, Ponani Gopalakrishnan, Ramesh Gopinath,
Stéphane Maes, Lazaros Polymenakos and Martin Franz

Human Language Technologies, Computer Science Dept.,
IBM T. J. Watson Research Center, Yorktown Heights, NY.,
email:rameshg@watson.ibm.com, phone: (914)-945-2794

ABSTRACT

This paper gives an overview of the IBM Large Vocabulary Continuous Speech Recognition system used in the 1996 Hub4 evaluation. It describes the acoustic and language models and adaptation techniques in the partitioned and unpartitioned evaluations. Evaluation results, analysis and further experiments are reported.

1. INTRODUCTION

Significant advances in speech recognition technology have been achieved recently, as seen on tests conducted with read speech corpora such as the Wall Street Journal corpus [1]. The focus of research has shifted recently to transcription of "found" speech like radio/TV broadcast news. Transcription of broadcast news presents technical challenges arising from several sources of signal variability. A typical broadcast news segment contains speech and non-speech data from several sources, such as the signature tune of the show, interviews with people on location - possibly under very noisy conditions - and interviews over the telephone, commercials, etc. Broadly speaking, the data in such broadcasts can be characterized using three criteria: the quality of the microphone or channel, the characteristics of the speaker, and the condition of the background. The signal might be acquired using a high quality microphone, a low bandwidth microphone, or could be telephone quality. The speaker may be an experienced announcer or correspondent or an inexperienced speaker.The speech from the former appears similar to read speech, whereas the latter produces largely spontaneous speech. The background may contain music, noise, or other interfering speech. In some cases, there is no speech present - the signal might consist of a musical interlude or an extended period of noise such as street noises added to evoke an environment.

Preliminary ideas to counter these problems were explored in the IBM system used in the ARPA sponsored, November 1995 Hub4 radio broadcast news transcription task [5, 7, 8]. The basic problem there was to transcribe an entire radio broadcast news show. This task has evolved into two tasks in the 1996 Hub4 evaluation - partitioned and unpartitioned evaluation. The unpartitioned evaluation is similar to the 1995 evaluation task - except that the test data is from both TV and radio shows. In the partitioned evaluation the test data is segmented into six categories (F0-prepared, F1-spontaneous, F2-degraded acoustics, F3-music background, F4-noise background F5-non-native speakers, and FX-other speech). The partitioned evaluation allows development of condition-specific systems since the data is pre-segmented. In the following sections we describe the overall system and the specifics of particular systems used in the partitioned and unpartitioned evaluations respectively. We also describe some experiments conducted after the evaluation.

2. SYSTEM OVERVIEW

The basic philosophy is to first try and identify the segments of input data that belong to one of several classes and use separate modeling techniques appropriate for each class. For instance, for the unpartitioned evaluation, segments detected as pure music are discarded and not decoded, segments identified as telephone quality speech are decoded by a system trained on telephone bandwidth speech, and so on. In the following sections, we describe techniques to handle issues in each class.

A brief description of our base recognition system follows (see [2, 4, 3] for details). The system uses acoustic models for sub-phonetic units with context-dependent tying. The instances of context dependent sub-phone classes are identified by growing a decision tree from the available training data [2] and specifying the terminal nodes of the tree as the relevant instances of these classes. The acoustic feature vectors that characterize the training data at the leaves are modeled by a mixture of Gaussian pdf's, with diagonal covariance matrices. The HMM used to model each leaf is a simple 1-state model, with a self-loop and a forward transition.

The acoustic training data used for the models in this paper comes from the following sources: WSJ-SI284 [5], MP-10 [5], BN-87 (the official 1996 Hub4 evaluation training data distribution consisting of 87 broadcast shows from radio and TV - about 30 hours of speech). The language model training data comes from the following sources: LM-BNA-96 and LM-BNA-95 (1996 and 1995 language model broadcast news archive texts distributed by LDC).

3. ACOUSTIC MODELS

There are about 5741 context-dependent phonetic states in our HMMs (that were originally built from WSJ-SI-284 data). For the 1995 Hub4 evaluation three models M95c, M95t, M95m (for clean speech, telephone-bandwidth speech and music-corrupted speech respectively) were built using MAP adaptation [10] on base models built on WSJ-SI284 for the 1994 Hub1 evaluation [5, 7].

3.1. PE models

For the partitioned evaluation (PE) separate models are built for each of the focus conditions F0 through F5. The F2 focus condition, however, contains a mix of telephone-bandwidth (BL) and full-bandwidth (NBL) speech data. On the development test data it was found that by separating the data into telephone and full-bandwidth speech and decoding separately with appropriate models the error rate reduced from about 57% to 43%. Therefore, separate models were built for the F2.BL and F2.NBL portions of this data. Models M96F0, M96F1, M96F4 and M96F5 were built using MAP adaptation of M95c models. For F3, M96F3 models were built in a multi-step process. First pure music segments from BN-87 were digitally added to WSJ-SI284 data. The model built from this data was MAP adapted to the F3 focus condition. For F2.NBL, M95c was MAP adapted using both the F4 (noise condition) and F2.NBL data to give M96F2.NBL. M96F2.BL was built by map adaptation of M96t models on F2.BL data.

3.2. UE Models

For the unpartitioned evaluation (UE) the PE models could not be used directly since the automatic segmenter (described below) could not be made to distinguish between these seven conditions. For example, the automatic segmenter could not distinguish between F0 (prepared), F1 (spontaneous) and F5 (non-native) speech accurately. However, the segmenter is reasonably accurate on separating music-corrupted speech (F3) and telephone-bandwidth speech (F2.BL). Telephone segments and music-corrupted segments could therefore be decoded with M96F2.BL and M96F3 models respectively. For clean data from the segmenter, M96F0+1 model was built by MAP adaptation on F0 and F1 data. For the remaining data, a “conglomerate” model, M96ALL using all training data from BN-87 except for conditions F2, F3 and FX. This models was built by reestimating the Gaussians from the BN-87 data alone without making use of MP-10 and WSJ-SI284 data.

4. LANGUAGE MODEL

We used a mixture model generated from the following sources: a) LM-BNA-96 - a deleted interpolation (DI) trigram LM and a maximum entropy (ME) trigram LM with class constraints were generated. b) LM-BNA-1995 (except for data from shows excluded by the Hub4 evaluation specification) - an ME trigram model was generated. c) a subset of LM-BNA-96 covering the time period 4/96-6/96 to capture recently - an ME trigram LM was generated. The recognition lexicon was selected from LM-BNA-95/96 It includes hyphenated words and is biased towards the 1996 data and towards non-hyphenated words. The lexicon size is 65185 words. The OOV rate on the eval data was about .7% and the perplexity of the LM was 172.

5. UNSUPERVISED ADAPTATION

For unsupervised adaptation on the test data three schemes were used. The first is iterative MLLR, the second is clustered transformations (CT) adaptation that was used on the clean data in our 1995 Hub4 evaluation [7], and the third is adaptation by correlation [12]. In our experiments CT adaptation worked best on clean speech provided there is enough test data (30s or more). Multiple iterations of MLLR was found to be marginally better than a single iteration on the development test data - and hence is used on segments that are of moderate length (> 10s). For segments less than 1s no unsupervised adaptation is performed while for short segments (between 1s and 10s) one iteration of MLLR is performed. ABC adaptation is a complementary scheme to MLLR that exploits the correlation between HMM states that is trained from a large training corpus. This scheme was tried only on F0 and F4 focus conditions on the development test data and gave moderate improvement over iterative MLLR and hence was used in the evaluation.

6. PARTITIONED EVALUATION

Initially, the test speech data is segmented into the seven classes (F0-F5 and FX) according to the markings in the pem file distributed by NIST. The F2 segments are further classified automatically into F2.BL (bandlimited) and F2.NBL (non-bandlimited). This classification (F2.BL/F2.NBL) is done using an automatic classification algorithm described in section 7.2. Using the same technique bandlimited portion of FX segments (FX.BL) are identified. The remaining FX segments are classified into categories based on the information in the distributed pem file. This is done in a hierarchical manner as follows: First segments with Low or Medium Fidelity are characterized as non-Bandlimited (FX.NBL - to be decoded like F2.NBL data) Then segments with background music are extracted (FX.F3 - to be decoded like F3 data), followed by segments that contain some other background (noise or speech) (FX.F4 - to be decoded like F4 data). Finally,

	F0	F1	F2	F3	F4	F5	FX	Total
A	21.6	30.4	38.9	28.0	42.2	30.8	54.2	32.2
A'	21.6	29.9	38.2	28.0	42.1	30.8	53.7	31.8
B	23.1	30.9	43.0	29.1	43.5	33.1	58.8	33.9

Table 1. WER on Evaluation: A) Official Eval Scores A') Eval Scores after procedural errors fixed B) Score on baseline system - before unsupervised adaptation.

segments that have foreign dialect and prepared mode are extracted (FX.F5 - to be decoded like F5 data) and the remaining segments form the final FX category (FX.F1 - to be decoded like F1 data).

The seven models M96F0-1, M96F3-5, M96F2.BL, and M96F2.NBL are used in a first-pass to decode the seven categories of segments in the test data obtained as above. The corresponding word error rates (WER) are shown in row B in Table 2. This is followed by unsupervised adaptation passes using the decoded scripts in the first-pass. Data corresponding to F1, F2.BL, F2.NBL, F3, and F5 are adapted using iterative MLLR (number of iterations depending on the amount of data as described earlier). The models for F0 are further auto-adapted using the ABC technique. An exception is long F0 segments ($>$ 30s) where only CT adaptation is applied. For F4 an automatic clustering algorithm is used to merge segments into classes so as to provide enough adaptation data for MLLR. This tends to merge segments with similar SNR. Iterative MLLR is applied on these segment clusters followed by ABC adaptation. The resulting decoded scripts were submitted for the partitioned evaluation task. The NIST official WER is shown in row A in Table 1. A procedural error in our submission was noted after the submission and the WER after accounting for this is shown in row A'. Notice that unsupervised adaptation helped in all conditions (even though the error rates in some conditions are nearly 50%). The gains from multiple iterations of MLLR and ABC adaptation on the actual evaluation were marginal. This was partly due to ABC adaptation parameters not being fine-tuned [12].

6.1. Additional Experiments

After the evaluation a few more models have been built leading to improvements in the base recognition performance. The first step was to built a conglomerate model that uses all the Hub4 training data (both MP-10 and BN-87) - M96H4. This model was further adapted using MLLR (the adaptation data for this was the condition-specific training data in BN-87 and MP-10) to generate the following models: M96H4F0, M96H4F1, M96H4F2.NBL, M96H4F4, and M96H4F5. As for telephone-bandwidth data, both MP-10 and BN-87 was bandlimited to build a conglomerate system M96TH4 (T-telephone). This model was MLLR-adapted on F2.BL data to give M96TH4F2.BL.

	F0	F1	F2	F3	F4	F5	FX	Total
C	22.9	30.8	45.0	32.7	40.1	32.1	57.1	33.7
D	22.9	30.8	38.2	32.7	40.1	32.1	57.1	33.1
E	22.9	31.0	37.0	31.0	37.9	26.8	57.1	32.7
F	21.8	29.8	32.7	27.9	39.9	25.8	53.8	31.1
G	22.5	29.5	37.2	28.4	39.2	30.1	57.1	32.1

Table 2. WER on Additional Experiments Results: C) Baseline score using M96H4 models D) same as C M96TH4 models telephone data E) MLLR to each condition from M96H4/M96TH4 F) unsupervised adaptation on E G) same as E with more transforms in supervised MLLR adaptation

The evaluation test data was decoded using this single conglomerate model, M96H4, and the results are shown in row C (comparable to row B). Interestingly, this single model (that only uses MP-10 and BN-87) comparely favorably with the seven models used in the evaluation. Notice however, that for conditions F2 and F3 the results are substantially worse than when condition specific models (M96F3, M96F2.BL, and M96F2.NBL) are used instead. Perhaps this is because there isn't sufficient training data for F2 and F3 conditions (relative to other conditions from which they are substantially different). To overcome this telephone segments (F2.BL and FX.BL) were decoded instead using the conglomerate telephone models (M96TH4) and the results are shown in row D. After (supervised) MLLR adaptation of these conglomerate model to each specific conditions (i.e., using models M96H4F0-1, M96H4F2.NBL , M96F3-5 and M96TH4F2.BL) the WER is shown in row E. The results seem to suggest that models for music-corrupted speech and telephone-speech have to be different from that for other types of speech. Row F corresponds to results after unsupervised adaptation of models for row E. Row G uses similar models to row E except that the number of MLLR transforms for supervised adaptation has been increased. Notice that F0 and F1 improved because of the increase in the number of transforms while F4 became worse. This is because there is not sufficient data to estimate the large number of transforms in the latter case. When there is enough data (as in F0 and F1) the increase number of transforms helps.

7. UNPARTITIONED EVALUATION

The system used for the unpartitioned evaluation is inspired by the approach that we followed for the HUB-4'95 evaluation [5, 7]. The segmentation algorithm was redesigned to segment the HUB-4'96 development data. Because of the limitations of the segmentation algorithm in separating pure speech (clean or prepared, spontaneous and foreign) from low noise conditions, all these conditions were pooled together and decoded with a conglomerate system described later on.

When applied to the evaluation data, the segmentation algorithm made some significant mistakes. A simplified approach is presented in this paper. The resulting word error rate on the NPR-Market Place is comparable to our 1995 results.

7.1. System description

The speech signal is automatically segmented according to the channel and background condition. A single pass decoder using rank-based decoding and envelope search (stack decoder) is used to produce an initial hypothesized script on each of the segments [2, 3]. The grammar and lexicon are the same as for the partitioned evaluation. The decoded word strings are used to seed an unsupervised iterative MLLR adaptation [9].

Depending on the duration of each segment, we use a different number iterations of the unsupervised MLLR adaptation procedure: no adaptation on segments shorter than one second, one-pass adaptation on longer segments which have a duration smaller than 10 seconds and three-pass adaptation on longer segments.

7.2. Segmentation

First, the distribution of feature vectors for each condition is modeled as a Gaussian mixture [5] trained on corresponding BN-87 data. For each feature vector x_t, and model M_j for condition j, $P(x_t/M_j)$ gives the likelihood of the frame coming from j. Since the condition is typically stable for a duration of a second or so, one imposes a minimum-length constraint on the segments. This is done by assuming a hidden Markov model for the generation of the input data as shown in Fig. 1. The j^{th} path in the model corresponds to the input data belonging to the j^{th} class, and the probability distribution of the arcs $c_{j,1} - c_{j,N}$ is given by M_j. The minimum length constraints on the segments are imposed by constraining the minimum length of the paths. The Viterbi algorithm is used to trace a path through the trellis corresponding to the model, and to assign a class id to contiguous sets of the input feature vectors. The model associated to each condition is called the signature of this condition.

The signatures (essentially Gaussian-mixture models for identifying the conditions) were built with 72 dimensional feature vectors (24-dimensional cepstra augmented with their first and second differences). The segmentation is achieved in four steps.

- Extraction of clean speech segments using signatures trained on MP-10.
- Segmentation of the data into the categories: bandlimited telephone, pure music, music/noise corrupted speech, speech. The signatures are trained on BN-87.

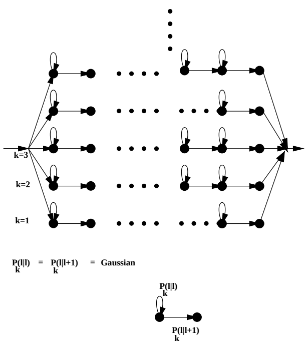

Figure 1. HMM models used for the condition signatures. Each branch k corresponds to a different condition.

- re-segmentation of pure music and music/noise corrupted speech into pure music, speech corrupted by music and speech corrupted by noise. The signatures are trained on BN-87.
- Bandlimited vs. non-bandlimited classification of music and speech corrupted by music segments, using signatures trained on MP-10 (telephone signals versus non-telephone signal).

We introduced signatures trained on MP-10 at the first and fourth steps. Indeed, on MP-10, the data tagged as telephone presented a more consistent behavior than the data tagged as F2 on BN-87. Also, BN-87 data tagged as F0 appeared more distorted than prepared speech from MP-10.

The fourth step results from the observation on '95 and '96 development data that the first three steps of the segmentation tag some telephone segments as pure music or music plus speech. On '96 development data, this strategy takes care of all the observed misclassifications of long telephone segments without reintroducing music or speech plus music into the telephone category.

The fourth step was also used on the F2 and FX conditions in the partitioned evaluation.

7.3. Acoustic models

The models used to decode each category were:

- The system M96F0F1 was used to decode the speech category (prepared, spontaneous and foreign).
- The conglomerate system M96ALL was used to decode the speech corrupted by noise categories.
- The telephone segments were decoded using M96F2.BL.
- The speech corrupted by music segments were decoded using M96F3.

7.4. Evaluation results.

Tables 4 and 3 present the results obtained on the evaluation data. The unsupervised iterative MLLR row presents the global evaluation results. File 1 to File 4 correspond respectively to shows CNN Morning News, CSP Washington Journal, NPR The World and NPR Market Place.

WER	Overall	File 1	File 2	File 3	File 4
Unsupervised Iterative MLLR	37.5	38.9	336.0	42.3	32.8

Table 3. Evaluation results corrected by adding scripts missing in our submission.

WER	F0	F1	Tele	F3	F4	F5	FX
Unsup. Iterative MLLR	26.0	34.1	41.4	53.7	42.8	34.1	60.3
File 1	34.7	32.6	0.0	81.3	43.3	0.0	47.1
File 2	22.3	36.2	40.5	0.0	28.7	0.0	0.0
File 3	26.0	24.5	0.0	53.3	60.0	29.3	58.0
File 4	20.5	37.2	50.7	45.9	45.7	35.7	97.2

Table 4. Evaluation results corrected by adding scripts missing in our submission.

The error analysis revealed that our segmentation introduced some mistakes not observed on the development data:

- The fourth step of the segmentation strategy introduced pure music and some speech corrupted by music segments into the telephone (BL) category.
- Some speech corrupted by music and speech corrupted by noise segments were categorized as pure music, non-processed by the decoder.
- Significant misclassification between the speech corrupted by noise and speech corrupted by music categories.

7.5. New system description

In order to improve the results a new version of M96ALL was built and run over the unpartitioned evaluation data - M96H4. The acoustic models have been rebuilt in order to improve the performances within each category and in order to increase the robustness of the decoding with respect to segmentation errors.

The segmentation algorithm was simplified to reflect the use of the new acoustic models. Also, the segmentation was rather trained on the carefully hand-labeled MP-10.

Again the segmentation separates the shows into different categories and category-dependent models are used for decoding.

The same lexicon and grammar were used. We also applied the same adaptation strategy.

7.6. New segmentation strategy

The new segmentation strategy extraction phases as illustrated in figure 2:

- Telephone extraction using signatures trained on MP-10.
- Pure music extraction using signatures trained on MP-10.
- The remaining segments constitute a conglomerate category.

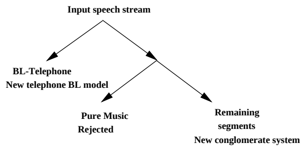

Figure 2. New segmentation strategy.

The resulting segmentation behaves significantly better on the evaluation data. However, a few telephone segments are still classified as pure music or conglomerate.

The same segmentation strategy was also implemented with signatures similarly trained on BN-87. This segmentation recovered more telephone segments, but lost segments of speech corrupted by noise or music which were classified as pure music. Also some pure music segments were introduced in the conglomerate. In conclusion, BN-87 training data leads to corrupted signatures models which severely degrade the segmentation accuracy. We hypothesize that differences in accuracy of the data labeling is responsible for the poorer accuracy.

7.7. Acoustic models

Based on the partitioned results and the segmentation mistakes, we have tried to limit the amount of category-dependent models and improve the performances of these models on their respective conditions. Therefore, we decided to use only two models:

- The conglomerate model M96H4 was used to decode the conglomerate category.
- M96TH4F2.BL was used on telephone segments.

7.8. Additional Experiments

Tables 5 to 7 summarizes the new word error rates. The new baseline row illustrate the WER obtained with the new models and the new segmentation without using the iterative MLLR adaptation.

WER	F0	F1	Tele	F3	F4	F5	FX
New baseline	23.5	33.2	43.8	33.9	44.0	39.8	63.9
Unsupervised Iterative MLLR	22.5	31.8	41.0	31.5	41.8	38.8	61.0

Table 5. Additional Experimental Results.

WER	Overall	File 1	File 2	File 3	File 4
Evaluation	37.5	38.8	36.0	42.3	32.8
Baseline	35.9	37.1	34.1	37.9	27.9
Unsupervised Iterative MLLR	34.2	37.1	33.7	37.9	27.8

Table 6. Results show by show.

The combination of a stabler segmentation algorithm.using non-corrupted signatures, with conglomerate Gaussians for speech decoding brings the WER roughly 3.1% absolute above the WER of the partitioned evaluation. This is in agreement with our observations on HUB-4'95 evaluation. Also the results on File 4 (i.e. NPR MarketPLace) is comparable to the overall results that we obtained in 1995 over similar evaluation data.

Acknowledgment

The authors would like to thank Dimitri Kanevsky in the speech department at IBM Watson Research for their help.

REFERENCES

[1] Proceedings of ARPA Speech and Natural Language Workshop, 1995, Morgan Kaufman Publishers.

[2] L. R. Bahl et al., "Robust Methods for using Context-Dependent features and models in a continuous speech recognizer", Proc. ICASSP, 1994.

[3] P. S. Gopalakrishnan, L. R. Bahl, R. Mercer, "A tree search strategy for large vocabulary continuous speech recognition", Proceedings of the ICASSP, pp , 1995.

[4] L. R. Bahl et al., "Performance of the IBM large vocabulary continuous speech recognition system on the ARPA Wall Street Journal task", Proceedings of the ICASSP, pp 41-44, 1995.

[5] P. S. Gopalakrishnan, R. Gopinath, S. Maes, M. Padmanabhan, L. Polymenakos, H. Printz, M. Franz, "Transcription of Radio Broadcast News with the IBM Large Vocabulary Speech Recognition System," Proc. of ARPA SLT Workshop, Feb 1996.

[6] L.R. Bahl, S. Balakrishnan-Aiyer, M. Franz, P.S. Gopalakrishnan, R. Gopinath, M. Novak, M. Padmanabhan, S. Roukos, "The IBM Large Vocabulary Continuous Speech Recognition System for the ARPA NAB News Task", Proc. of ARPA SLT Workshop, Jan 1995.

[7] P. S. Gopalakrishnan, R. Gopinath, S. Maes, M. Padmanabhan, L. Polymenakos, "Acoustic Models Used in the IBM System for the ARPA Hub 4 Task," Proc. of ARPA SLT Workshop, Feb 1996.

[8] L. Polymenakos, M. Padmanabhan, D. Nahamoo, P.S. Gopalakrishnan, "Suppressing background music from music corrupted data of the ARPA Hub 4 task," Proc. of ARPA SLT Workshop, Feb 1996.

[9] C. J. Legetter and P. C. Woodland, "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous density HMM's", Computer Speech and Language, vol. 9, no. 2, pp 171-186.

[10] J. L. Gauvain and C. H. Lee, "Maximum-a-Posteriori estimation for multivariate Gaussian observations of Markov chains", IEEE Trans. Speech and Audio Processing, vol. 2, no. 2, pp 291-298, Apr 1994.

[11] M. Padmanabhan, L. R. Bahl, D. Nahamoo, M. A. Picheny, "Speaker Clustering and Transformation for Speaker Adaptation in Large Vocabulary Speech Recognition Systems", ICASSP-96, vol. II, pp 701-704.

[12] S. Chen, "Adaptation by Correlation" (elsewhere in this proceedings).

WER	F0	F1	Tele	F3	F4	F5	FX
File 1	34.4	32.0	0.0	44.4	43.5	0.0	47.1
File 2	19.4	33.5	39.8	0.0	24.6	0.0	0.0
File 3	17.7	21.7	0.0	26.7	26.7	30.7	60.2
File 4	17.2	34.9	53.4	30.9	41.1	41.5	74.5

Table 7. Results show by show and condition by condition.